Corpus *Réfugié·es*

> ***"Je suis doctorante en sociologie, je débute une thèse sur les réfugié·es à travers le monde au 21e siècle. Je souhaite construire une bibliographie sur le sujet, distinguer la part de documents traitant des réfugié·es climatiques et des réfugié·es politiques et enfin identifier leur origine géographique."***

1. Description générale

Pour répondre à cette problématique, il faut constituer un corpus de documents traitant des réfugié·es. Istex, qui est un réservoir d'archives scientifiques mondiales, est une ressource particulièrement pertinente. Ce corpus va ensuite être exploré grâce à l'outil Lodex pour me fournir à la fois les informations qui m'intéressent et les ressources documentaires les plus judicieuses vis-à-vis de ma recherche.

- **Objectif** : repérer les études sur les réfugié·es.
- **Outil de TDM utilisé** : Lodex et les web services associés.
- **Contraintes imposées par l'outil** : Istex propose un format de sortie adaptée à l'outil Lodex. Pour utiliser les web services qui nous intéressent, il faut vérifier que les documents sont en anglais et contiennent des résumés.

Exercice 1 – Construire une requête Istex

Étape 1 : Se rendre sur Istex Search : https://search.istex.fr.

Étape 2 : Rechercher les formes anglaises et françaises *refugee, réfugié, asylum seeker* et *demandeur d'asile.*

- Pour une aide sur la syntaxe des requêtes consulter la documentation : https://doc.istex.fr/tdm/requetage/.
- Les espaces blancs sont considérés comme des opérateurs booléens OR (sauf en recherche assistée).
- Par défaut, la requête est insensible à la casse.
- Les guillemets permettent des recherches exactes (notamment avec des espaces). Ils sont inutiles en recherche avancée.

Étape 3 : Limiter le bruit et le silence.

- Rechercher les variantes (singulier/pluriel et féminin/masculin) de chacun des termes de la requête.

- Par défaut, la recherche s'effectue dans tous les champs interrogeables d'Istex. Pour affiner la recherche, on peut préciser les champs les plus à même de renvoyer des résultats pertinents (soit le titre, le résumé et les mots-clés d'auteur dont la dénomination est *title*, *abstract* et *subject.value*). La requête prend alors la forme *champ:()*. La liste des champs est accessible dans l'assistant à la construction de requête.
- Pour répondre aux contraintes imposées par l'outil, il faut s'assurer de la présence de résumés dans les documents et s'assurer qu'ils sont en anglais.
- Pour répondre à la problématique, il faut limiter les dates de publication au 21e siècle.
- Pour limiter le bruit, on doit s'assurer de ne sélectionner que des articles de recherche.

Étape 4 : Pour aller plus loin… transformer l'équation en utilisant des expressions régulières.

Les expressions régulières, présentées entre barres obliques (/), vous permettent de raccourcir votre requête.

- les . signifient n'importe quel caractère ;
- les * signifient n'importe quel nombre fois (s'appliquent au caractère précédent) ;
- les ? signifient entre 0 et 1 fois (s'appliquent au caractère précédent) ;
- les crochets [] sont l'équivalent d'un OR ;
- les .raw considèrent que le token interrogé est le champ dans sa globalité.

Pour en savoir plus et tester des expressions régulières se rendre sur https://regex101.com/.

Exercice 2 – Premier pas vers le TDM

Étape 1 : Télécharger le corpus.

- Extraire le corpus *Réfugié·es* en utilisant l'équation définie et en choisissant le format adapté pour un import dans Lodex.

Étape 2 : Importer le corpus dans Lodex.

- Se rendre sur votre instance Lodex : se connecter avec votre nom d'utilisateur et votre mot de passe.
- Aller dans l'interface administrateur en cliquant sur "Voir plus > Admin".
- Importer le corpus en glissant le fichier *.zip* que vous venez de télécharger **sans décompression préalable**.
- Choisir un loader. Un loader est un script d'adaptation du fichier à Lodex. Il dépend du format de fichier fourni en entrée. Choisissez le loader "ZIP résultat de dl.istex.fr". Cliquer sur "Importer les données".

- Pour rappel, un modèle est un fichier *.tar* qui permet de mettre en forme le site créé avec Lodex. Importer le modèle fourni , cliquer sur le menu en haut à droite “Modèle > Importer un modèle”.
- Publier votre site pour une première visualisation en cliquant sur “Publier” en haut à droite. Cliquer sur l'icône en forme d'œil pour voir le résultat. Explorer les différents graphiques à partir de l’onglet “Graphiques” en bas. Consulter quelques ressources grâce à l’onglet “Recherche”pour vérifier que tout fonctionne.

Étape 3 : Extraction de mots-clés des résumés via le web service Teeft.

- Aller dans “Données > Enrichissements > + Ajouter”, puis donner le nom “Mots-clés (WS)”, aller chercher le web service **Teeft** approprié dans le catalogue (bouton vert à droite du champ “URL du web service”). Choisir “Résumé” dans la colonne de la source, cliquer sur “Sauvegarder”. Cliquer enfin sur “Lancer”.
- Dans Lodex, avant de faire un graphique, il est nécessaire de déclarer la colonne comme une ressource : dans “Affichage > Ressource principale”, créer une ressource “Mots-clés” à partir de l’enrichissement “Mots-clés (WS)”. On notera que l’**identifiant de la ressource** est une suite alphanumérique (sensible à la casse) de 4 caractères. L’identifiant est inscrit entre crochets à côté du nom de la ressource.
- Aller dans “Affichage > graphiques > + Nouveau champ”, puis créer le graphique “Mots-clés les plus représentés dans les résumés” à partir de la ressource “Mots-clés Teeft”. Choisir la routine “distinct-by” (dans général) puis ajouter derrière l’URL **l’identifiant de la ressource** “Mots-clés Teeft”. Enfin, dans “Affichage”, choisir le format “Diagramme en barres” en filtrant les résultats (mettre “valeur minimum à afficher” à 5).

Étape 4 : Extraction des thématiques du corpus.

- Dans “Données > Précalculs > + Ajouter”, lancer le précalcul *lda-segment* sur la colonne “Résumé”.
- Les précalculs peuvent directement être utilisés pour faire des graphiques. Aller dans “Affichage > graphiques > + Nouveau champ”, puis créer le graphique “Thématiques extraites du corpus” à partir du précalcul *lda-segment*. Choisir la routine “segments-precomputed-nofilter” (dans général). Enfin, dans “Affichage”, choisir le format “diagramme à barres groupées”.

Étape 5 : Création d’une carte en fonction des pays mentionnés dans les articles.

- Certains enrichissements sont déjà effectués sur l’entièreté de la base Istex. Pour obtenir des entités géographiques, aller dans “Données > Enrichissements > + Ajouter”, puis donner le nom “Entités nommées de géographie”, cliquer sur “Mode avancé” et copier le code ci-dessous, cliquer sur “Sauvegarder”. Cliquer enfin sur “Lancer”.

```
[assign]
path = value
value = get("value.Entités nommées
(Unitex).placeName").split(",").get(0)
```

Remarque : Le mode avancé offre une meilleure flexibilité pour transformer les données. Il n'est néanmoins pas nécessaire de comprendre le code. Les transformations les plus usuelles sont disponibles via ce lien.

- Aller dans "Données > Enrichissements > + Ajouter", puis donner le nom "Pays et subdivisions (WS)", en utilisant le web service "Associer un terme au vocabulaire Pays et Subdivisions" sur la colonne nouvellement créée "Entités nommées de géographie".
- Pour construire la carte, nous avons besoin du code pays à 3 lettres, retourné par le web service dans le champ "cartographyCode". Dans "Affichage > Ressource principale", créer une ressource "Code pays" à partir de l'enrichissement "Pays et subdivisions (WS)". Après avoir sélectionné la colonne, cliquer sur "Ajouter une opération" pour appliquer une transformation "GET" sur le path "cartographyCode".
- A l'aide de la documentation Lodex, créer la carte (graphique nommé "Cartographie"). Ne pas oublier de rajouter l'identifiant de la ressource "cartograhyCode" après la routine.

Pour aller plus loin : augmenter l'exhaustivité de la carte.

Pour des raisons d'efficacité, nous avons réduit la quantité des données lors de l'étape d'avant. Si il reste du temps, relancer les deux enrichissements précédents en modifiant l'enrichissement avancé "Entités nommées de géographie" par :

```
[assign]
path = value
value = get("value.Entités nommées
(Unitex).placeName").split(",").get(0)
```